Conceptualising Information Production in the Context of the SDHSS Ontology Ecosystem

Francesco Beretta

Methodos
Savoirs et textes

24 | 2024
Digital Humanities in the Web 3.0 Era

Conceptualising Information Production in the Context of the SDHSS Ontology Ecosystem

Conceptualiser la production d'information dans le cadre de l'écosystème d'ontologies SDHSS

Francesco Beretta

Electronic version
URL: https://journals.openedition.org/methodos/11110
DOI: 10.4000/12xqn
ISSN: 1769-7379

Publisher
Savoirs textes langage - UMR 8163

Provided by Bibliothèque Diderot de Lyon - ENS

Electronic reference
Francesco Beretta, "Conceptualising Information Production in the Context of the SDHSS Ontology Ecosystem", *Methodos* [Online], 24 | 2024, Online since 14 November 2024, connection on 27 December 2024. URL: http://journals.openedition.org/methodos/11110 ; DOI: https://doi.org/10.4000/12xqn

This text was automatically generated on December 21, 2024.

Conceptualising Information Production in the Context of the SDHSS Ontology Ecosystem

Conceptualiser la production d'information dans le cadre de l'écosystème d'ontologies SDHSS

Francesco Beretta

Introduction

1 While large language models (LLM) and AI-based audio-visual production technologies are currently attracting a great deal of public attention, from a humanities and social sciences (HSS) perspective we should not overlook the importance of the Semantic Web and Linked Open Data (LOD) which are revolutionising the way information is analysed and shared at scale[1]. As a matter of fact, if generative AI creates new representations of the world in the form of language, images and videos, with more or less hallucination, knowledge graphs aim to identify objects in the world and capture their properties and relationships as accurately as possible in a previously unimaginable scale. Google's Knowledge Graph, announced in 2012, has been defined by the company as a "giant virtual encyclopaedia of facts": "By March 2023, it had grown to 800 billion facts on 8 billion entities"[2]. By transforming unstructured text into structured data with explicit semantics, it is possible to improve the precision of search engine results and derive new information about the objects under consideration.

2 Given this context, will HSS researchers be able to share the information they produce daily by leveraging the same semantic technologies, and interconnecting their data with the authority files and metadata of libraries, archives and museums, in order to create a giant graph of high-quality information about the objects their discipline studies? And thus enable a renewed knowledge of past societies and a better understanding of present issues? In earlier work, I presented the methodology adopted

in the *Semantic Data for Humanities and Social Sciences* project (SDHSS) in order to develop an open and refinable shared conceptualisation that carefully documents the meaning of the published data and facilitates reuse for new research[3]. The most relevant and valuable content of research data produced by HSS is *information* defined as a representation of objects, their properties and relationships. This is in line with the standard definition of ontology in the context of the Semantic Web, formulated by Tom Gruber as "intended for modeling knowledge about individuals, their attributes, and their relationships to other individuals"[4]. If we apply established methods of ontological analysis in order to develop an ecosystem of shared and reusable ontologies, we will obtain a rich universe of reusable information. Such a universe will allow us to represent multiple facets of reality in a cumulative distributed graph of increasing volume and quality.

3 The online application OntoME, developed under my direction at the Laboratoire de recherche historique Rhône-Alpes (LARHRA) since 2017[5], has been designed as a support tool to facilitate the implementation of this vision, allowing different projects to adopt data models or application profiles specific to their research, while reusing existing ones as much as possible[6]. In this context, ontology is primarily understood, according to a standard definition, as a "set of representational primitives with which to model a domain of knowledge or discourse"[7], and not as a collection of instances or types. In other words, the SDHSS project's priority is to produce terminological knowledge (which can be expressed in RDFS or OWL-DL) and not assertional knowledge (which is expressed using RDF)[8]. Using this common ontology ecosystem, data can be produced in terms of instances of classes belonging to the shared conceptualisation in a variety of distributed, interconnected information systems, e.g. Geovistory[9], also referring to common authority files like IdRef or Wikidata[10].

4 It seems reasonable to adopt the CIDOC CRM to achieve the much-desired cross-disciplinary interoperability. This standardized conceptual model (ISO 21127:2014), aiming at the integration of cultural heritage information, can be used as a core ontology providing the high-level classes needed to describe factual, spatio-temporal information in the HSS domain. But it is essential, at the same time, to increment it with a high-level extension, in the SDHSS core namespace[11], in order to cover the entire domain of research of HSS and furthermore add subdomain extensions, at different levels of abstraction, that are needed for more specific research agendas. This coherent ecosystem of ontologies, in conjunction with the application profiles managed in OntoME (i.e. subsets of classes and properties pertinent to a specific research project), and utilised in production within distributed information systems, will provide HSS with a range of reusable conceptualisations. This type of conceptualisations ensures data interoperability that is much richer semantically than simply aligning heterogeneous ontologies, and much less costly in terms of time and resources than having to reinvent a conceptualisation for each project or sub-domain.

5 This promising vision for the renewal of research in the HSS[12] appears to be challenged by several objections. We can group them into three main categories. The first is related to the fundamental assumption of the SDHSS vision, i.e. the possibility of collaboratively developing a cross-disciplinary shared ontology ecosystem: if information production stems from different scientific disciplines, and is driven by different research agendas, isn't this a major and quasi-structural obstacle to the reuse of data? Is a representation of factual reality through information really possible, or at

least expressible in the form of interoperable data? What does it mean to talk about facts or objectivity in the context of HSS data production? Google does it, calling its knowledge graph an "encyclopaedia of facts", but is this allowed in the field of HSS research that is essentially driven by constructivist approaches[13]?

6 The second category of objections relates to the question of the identity of the intended objects. On the one hand, can we be sure that a person mentioned in an ancient or medieval document is the same as a person already identified in an information system? Or are these in fact two different persons, considering different events known about them? And what about the concept of astronomy, is it the same in the 16th and the 21st century, given the completely different understanding of this scientific discipline in the two historical periods? On the other hand, even if we know we're considering the same event, like World War II, is it really the same object if it is conceptualised in different ways, and are we allowed to express an identity relationship using the *owl:sameAs* property? Wikidata considers it as being an object of type 'world war'[14], and this class appears to be, through inheritance, a subclass of 'occurrence', which is an 'occurrence of a fact or object in space-time', also known as an 'event'[15], whereas the related French National Library RAMEAU authority file considers this war as a term or topic, i.e. a concept[16]. Given this difference in ontological substance, do the respective identifiers or URIs refer to the 'same thing'[17]? And what about the '2023 Israel–Hamas War'[18], also identified in the Google Knowledge Graph[19]: can it be considered the same event even if interpreted in very contrasting ways by different parties? Is objectivity possible at all in relation to events of this kind, or are different sides talking about different facts[20]?

7 Finally, there is the challenge of uncertainty, in its various forms, whether caused by a lack of information or by inconsistencies or contradictions about the same facts in different documents: which one is the correct birth year of a person? Or the institution the person belonged to, but which is known under different names? This challenge is often related to the issue of information sourcing, quality and reliability but it can also concern the question of identity: were there actually one or two different persons, with different birth years, or two institutions with different names? These challenges raise issues that are at once epistemological, semantic and methodological[21].

8 This paper is intended as a contribution to overcoming these challenges. It conceptualises the information production process applying several methodological principles developed by scholars active in these different fields that serve as the foundation of the SDHSS project's ontology ecosystem. Furthermore, the methodology is informed by my 15 years of experience in data modelling and management in several research projects conducted by PhD students or in funded collaborative projects. This experience was initially gained in the context of the *symogih.org* project, and subsequently in Geovistory, a collaborative VRE aimed at creating a collectively managed knowledge graph based on the semantics of the SDHSS ontology ecosystem and the application profiles managed in OntoME[22].

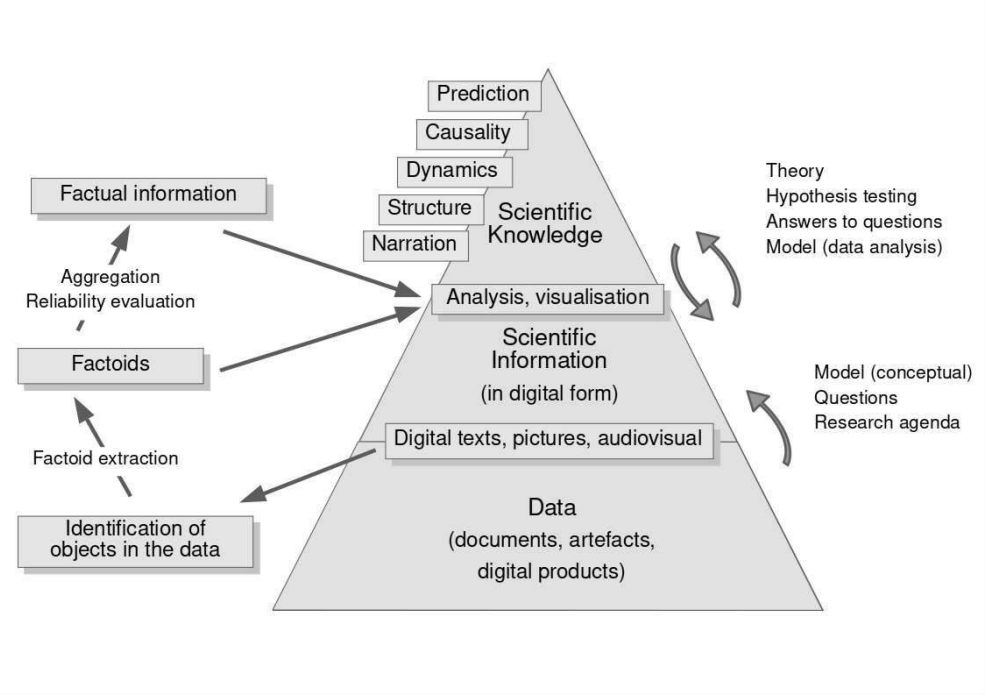
9 First, I will analyse the different epistemic levels of information and show how they articulate with the domain of data and knowledge. Indeed, an important distinction must be made in HSS between information and knowledge: information can be defined as a representation of observed objects, their properties and relationships, whereas knowledge appears as an interpretation of reality, a model expressing the researchers' understanding of complex phenomena, their causes and likely evolution. From this

perspective, the so-called ‘knowledge graph’ should be actually called an ‘information graph’ because knowledge resides in the minds of researchers, or in the texts they publish to share it. I will then present some of the relevant principles of ontology engineering established by the OntoClean methodology and used in the DOLCE and DnS foundational ontologies to produce a high-level model of human discourse about scientific objects. On this foundation, we will analyse the CIDOC CRM and discuss the main classes of the SDHSS extension, which can ground a cross-disciplinary conceptualisation of the HSS domain and serve as a basis for the conceptualisation of the information production process. Finally, a detailed analysis of specific aspects of this process will be conducted, focusing on information extraction from texts and the aggregation of existing semantic graphs. The epistemological and ontological principles outlined will be employed to generate a conceptualisation of the process of information production that responds to the objections discussed above.

Epistemic Levels of Digital Data in HSS

10 The aim of HSS disciplines is the production of knowledge about human beings, as individuals or as collectives, their characteristics, activities and social life. This process of knowledge production can be represented by a re-interpretation of the “data, information, knowledge, wisdom pyramid” (DIKW) developed by information sciences[23], which is very helpful in highlighting the different epistemic levels in line with our epistemological analysis (Fig. 1). I will present here an approach that is based on my experience as a historian with a social science perspective, but which, in broad outline and with the necessary adaptations, is applicable to other fields of HSS research interested in collecting complex or extensive information about human individuals and societies.

Fig. 1- Pyramid “data, information, knowledge” as interpreted in HSS



11 All research must begin by defining a research agenda that fits within the horizon of existing knowledge as expressed in the literature, and that defines the methodology to be used and the general questions to be answered, or theories to be established or challenged[24]. More specific research questions will also be formulated in this first phase which define the lines of inquiry and are essential in order to be able to choose the documentation to be used or the surveys to be carried out, and to define which information will be gathered.

12 At this stage we are at the base of the pyramid, at the level of *data*, a term we use here in its primary and etymological sense, derived from the Latin *datum*, i.e. everything that is regarded as given and perceived independently on the observer, and not in the sense of digital data. By data, we mean the artefacts and other traces of the observed human reality that are given as such, whether directly observable in the social sciences or indirectly - through sources and physical remains - in the historical sciences. The data defined in this way are largely analogue, but in recent decades they are increasingly present in digital media. These data, as original traces of the human and social world, must not be confused with their digital representations in the form of digital transcriptions, photographs, audiovisual recordings, etc. Data, in the sense of digital representations, constitute the first level of scientific information intended as a representation of the objects of the world. These digital artefacts (pictures, etc.) are information in the sense we use it here, but unlike sketches of objects on paper or printed editions of texts, they are produced in digital form and stored in information systems: digital data are therefore carriers of scientific information that represent the original data, the traces of human activity, or what these allow us to know about human activity itself (Fig.1). Distinguishing between these two substantially different types of data, the material and the digital, often confused in the literature, is important.

13 Based on their research questions, HSS researchers must make a selection from the mass of sources, or any other available and/or experimentally constructed traces of

human activity, i.e. the data in the original sense, in order to extract from them the information that will be analysed and serve as a basis for knowledge production. At this stage the methods for producing information in digital form must be chosen. Digital images, raw text, and unannotated video material are considered *unstructured data* because they carry no explicit semantics. In contrast, if you choose to use a formalism such as the Text Encoding Initiative (TEI) and produce documents tagged in XML, or if you transcribe documents by populating spreadsheets, you introduce a model and produce *semi-structured data* with meaning expressed by XML tags or spreadsheet columns. The full potential of information as a representation of objects in the world emerges when you create *structured* data. Relational databases or graph technologies identify objects by database primary keys or URIs, formalizing the semantics of their properties and relationships in a more or less well-documented conceptual model. This identification allows for the expression of complex relationships in space and time between different types of objects in a way that is impossible with spreadsheets or XML-encoded text.

14 Information is at the heart of the scientific process and when expressed in the form of structured data, especially in the form of semantic graphs, it makes it possible to capture all the richness of the characteristics of the objects studied. At the same time, the challenges mentioned above appear precisely at this stage: how to adopt a cross-disciplinary conceptualisation that allows the reuse of scientific information for new research? How do we answer the question of the identity of the objects, the uncertainty of their identification, and the quality and reliability of the information collected in the form of digital data? The answer lies precisely in the fact that, in line with the principles of semantic technologies, we define information as a *representation* of objects in the world (people, organisations, artefacts, tokens in texts, etc.), their characteristics (physical properties of objects, education and income levels of people, opinions, etc.) and their relationships in time and space (membership in organisations, exchange of messages or goods, journeys, etc.). But "representations always simplify objective reality, literally 're-presenting' it"[25]. This means that information, even if conceived and produced with an explicit desire for objectivity, is always constructed, it is always the result of a question or a point of view and, in the context of digital information systems, it always involves, by definition, a certain degree of simplification (it is the result of applying a specific conceptual model) and approximation (depending on its sources and the quality thereof). The challenge, then, is precisely to reflect and document this conceptual and methodological *approximation*.

Factoids and Factual Information

15 A crucial element of this endeavour is the distinction that has to be made between two different epistemic levels of information expressed in the form of structured data: factoids and factual information. One can aim at a faithful reproduction of the content of different documents about the same facts, or at a detailed measurement of all the economic transactions in one day or of the more diverse, unverified and not categorized interactions between two social actors, situating oneself on an epistemic level that can be called that of *factoids.* In this scenario, we will have access to extensive but redundant, or even contradictory information about the same properties of objects. This information can be very relevant for some research agendas with focus on details

but, taken as such, it will inevitably distort the results of the analyses, at least in a research perspective that aims at a more general understanding of the functioning of human groups, the evolution of prices or the interaction of social actors. With this in view it is necessary to aggregate the available information in order to identify and reconstruct aspects or segments of the activity of actors, individually or in groups, so that it provides a more abstract and consistent representation of facts. In the event of disagreement between factoids, it will be necessary to make choices, generally based on the quality of the available documentation, so that the analysis is not distorted by the redundancy of the facts. The epistemic levels of factoids and factual information are therefore fundamentally different.

16 John Bradley and Michele Pasin have sought to formalise this distinction in an article that publishes a *factoid* data model, developed in the context of prosopography projects for the Middle Ages undertaken by the Department of Digital Humanities at King's College London. On one side, they explain, there are "*states of affairs*", on the other side is what the sources assert regarding these same facts: "The factoid approach prioritizes the sources, rather than our historians' reading of them"[26]. In other words, factoids express the content of the documents, whereas the epistemic level of factual information aims to represent individual and social characteristics as such, i.e. 'facts' as historians can reconstruct them. In order to go from one to the other level of information historians must apply the methods of criticism, such as conjecture, inference, contextualization, etc., with the aim of verifying the reliability and degree of veracity in each assertion made by the source, then aggregating the content of the various sources into *factual information* which is intended to be the best possible approximation to the factuality of the characteristics and relationships of the objects studied (Figure 1).

17 Three important points need to be made at this stage. First, the most relevant distinction between factoids and factual information is that of the epistemic level they express. This is true even if these different kinds of structured data share the same conceptualisation of the characteristics of human individuals and societies: it is their epistemic relation to the documents, and to the facts they refer to, that makes the difference, not the data model they imply. In other words, the same ontology can be used to express graphs of factoids or of factual information. However, it is important to be aware of the different epistemic levels of these two types of graphs.

18 Second, the epistemic level of factoids stems and can cover two different aspects of information production: on the one side, the aim to convey the precise content of one specific document; on the other side, the issue of uncertainty in the identification of individuals and of reconstructing their properties, and therefore the preference by researchers to store factoids in a distinct graph from the one of factual information.

19 Third, this epistemological analysis can be applied to two common use cases in information production. The first is information extraction from texts, whereby the produced graphs will naturally have the epistemic level of factoids. This is not only due to the difficulty of identifying objects mentioned by so called 'named entities', but also because the extracted information reproduces the point of view of its source. The second is aggregation of existing information graphs: from the perspective of researchers, the information published by other projects, or even by public authority files, even if they use the same ontology for expressing factual information, has the epistemic level of factoids. In order to meet the needs of one's own research agendas, a

process of information quality verification and data aggregation will be required. As a general rule, when data is submitted for analysis, it is essential to use information that is consistent, non-redundant and non-contradictory in representing the same state of affairs to avoid distortions in the results. The above considerations and distinctions will be important when discussing the information production process later.

20 Returning to the DIKW pyramid model, once the information is available in the form of factoids or factual information, it needs to be coded and simplified according to the lines of inquiry by, for example, reducing an excessively large number of modalities of a qualitative variable to a set of significant types or grouping a variety of quantitative values into a limited number of value classes. However, it is essential to maintain a clear distinction between this step and the previous one, which involves producing information intended to be an accurate and factual representation of objects and their characteristics. This is essential to enable scientific information reuse for new research.

21 At this stage, the research queries are applied to the collected information, available in the form of digitally encoded data, using different kinds of digital analysis and visualisation tools (statistical software, network analysis, spatial analysis, etc.) or machine learning methodologies. The model (in the statistical sense) that emerges from these analyses has an eminently heuristic function because the mathematical and visual representations produced by analysis software always require critical discussion, contextualisation and interpretation. At the same time, the analysis software and AI methods make it possible to make visible significant phenomena that would otherwise be impossible to see "with the naked eye", despite the considerable volume and complexity of the available information.

22 At the end of this process, researchers produce knowledge, answering the questions of their research agenda and publishing the results of their investigations. Knowledge can be produced in two ways. Either through an inductive process, using the information gathered to build a model of the human or social phenomena under consideration, or to develop a theory and account of complex social phenomena, providing an interpretation of their structures or causes. Or, through a deductive approach, empirically testing hypotheses derived from the theory on the basis of the information available[27]. The insights and theories that knowledge produces always involve a synthesis of information and an interpretation that goes beyond the simple representation of factual reality. This analysis clearly shows that there is an essential epistemic distinction between the knowledge produced in this process and the information on which it is based. Therefore, in the logic of open science, it is essential to publish not only the knowledge obtained but also the research data themselves, i.e. the information collected, in order to facilitate the verification of the hypotheses put forward by exposing them to "falsification" in the logic of a reproducible scientific approach[28].

Ontology Engineering and Epistemology: OntoClean and Scientific Objects

23 This raises the question of the methodology to be adopted in order to develop a conceptualisation capable of supporting the production of information that can be reused in a cross-disciplinary context. This is the task of ontology engineering, which "is concerned with making representational choices that capture the relevant

distinctions of a domain at the highest level of abstraction" and inherits from philosophical ontology "a rich body of theory about how to make ontological distinctions in a systematic and coherent manner"[29]. With this in mind, we'll present now some aspects of the OntoClean methodology, developed by Nicola Guarino and Emil Welty, as a "formal foundation of ontological analysis"[30] to recall some basic aspects of ontology engineering. According to this methodology, classes are sets of entities that share properties in common. In OntoClean, properties primarily define the meaning of classes and their *intension*, whereas a class and all its instances, i.e. the entities that share that property, represent the *extension* of the class. *Properties* are therefore used in OntoClean to define meaning, *intension,* and not relationships between instances as in RDF.

24 Properties define the *essence* of entities: they are *rigid*, if they are essential to all possible instances of a class during these instances' whole life; or *anti-rigid*, if they are merely accidental, time-limited or optional. If we take the example of a church, the property of being a building is essential to the entities of this class, and therefore *rigid*, while the property of being used for worship is not, because the building could at some point in its life be used as a stable or a disco. It must be noted that a class defined by a rigid property, like being a human, cannot be subsumed by an anti-rigid class, like being a student: "student" can only be a subclass of "human," and not vice versa. We can also observe that the property of being a student for a person, as well as being a church for a building, does not appear to be an intrinsic quality of the respective entities, but an external classification or accidental use, and therefore does not define their essence[31].

25 According to OntoClean, properties can also provide criteria to discuss the *identity* of entities: which properties or features allow us to say that these two buildings are identical? Applying this methodology, one could say that if these two entities occupy the same place in space, and there is no interruption in their physical existence, they are the same entity, even if their use or name changes. OntoClean distinguishes "between properties that carry an identity criterion and properties that do not". In the example of the church, the first is a physical quality, and the second is a temporal state or use. The same applies to the notion of *unity* where properties allow us to define if entities are wholes or parts of other entities. When we consider amounts of matter, such as water, they have no intrinsic criterion of unity. An ocean is a body of water whose unity is determined by the continents that enclose it.

26 *Essence*, *identity* and *unity* are basic notions in OntoClean that we must consider to carry out the foundational analysis of a domain conceptualisation in order to produce a robust ontology. Two key points need to be made at this stage. On the one hand, modelling experience shows that time-limited, anti-rigid properties often result from relationships with other entities or contexts, such as classifications that express a human association of a concept with an instance of a class. This approach makes it possible to distinguish between intrinsic properties of objects and external interpretations or social roles attributed to objects by humans or societies. On the other hand, as defined by intensional properties, essence is not substance in the Aristotelian sense. It is the expression of attributes observed in and abstracted from *things,* an abstraction that defines *scientific objects* in the context of a scientific discipline. We adopt here the epistemological analysis that has been developed by Evandro Agazzi, among others, which emphasises the difference between *things* in the

world and *objects* in human and scientific discourse. Scientific objects are defined by predicates that result from operations by empirical sciences that identify "attributes (that is, properties, relations and functions) that are present in things". These "structured sets of attributes" – the properties in OntoClean – make possible the "clipping" of scientific objects out of things in the world. Such a "clipping" occurs according to a scientific discipline's specific point of view.[32].

27 We need to be aware of this relevant distinction when analysing the formalised conceptualisations of classes and properties that make up an ontology, and that are aimed at representing "individuals, their attributes, and their relationships to other individuals", because this "representational vocabulary" is not about *things* as such, but it's much more an artefact defining "concepts, relationships, and other distinctions that are relevant for modelling a domain"[33] , thus "clipping" the *objects* of a scientific discipline or domain of discourse out of the *things* in the world. This epistemological approach allows us to understand that ontologies emerge from a commitment to a particular domain of discourse, in our case a scientific discipline and this can be an obstacle to straightforward interoperability because scientific objects are conceptualised from different perspectives and can have different meanings.

28 But it's precisely the vigilance that this epistemological approach requires, in conjunction with the OntoClean methodology, that makes it possible to ground a robust methodology for performing high-level semantic analysis of different ways of modelling the same 'things', in order to define a more abstract conceptualisation of the general domain of HSS research. We will therefore need to look for sufficiently generic and rigid intensional properties to define common classes that can be accepted in the context of the field of different HSS scientific disciplines, while allowing the latter to develop subsequent specialisations, defining scientific objects specific to their domain of discourse, as extensions of the high-level ones. This process is the prerequisite for achieving semantic cross-disciplinary interoperability[34].

Foundational Ontologies: DOLCE and DnS

29 In a critical and stimulating article on the subject, Giancarlo Guizzardi writes that information interoperability is only possible if we adopt "formal, shared and explicit representations of conceptualisations, or what the field of knowledge representation has conventionally called ontologies". This author specifies that what constitutes an ontology is not the fact of expressing the conceptual model of a particular project by means of formal logic or the *Ontology Web Language (OWL)*, but rather the fact of carrying out an analysis of the essential aspects of reality, such as the identity of the objects, their relations, their compositions and their dependencies, by adopting a high-level conceptualisation that is cross-disciplinary and can be applied to several fields of scientific discourse[35]. This is the role of foundational ontologies[36], a research field to which Guizzardi has contributed as one of the creators of the *Unified Foundational Ontology* (UFO).[37]

30 Among foundational ontologies, the *Descriptive Ontology for Linguistic and Cognitive Engineering* (DOLCE) is well adapted to the HSS perspective. It is an ontology of particulars, i.e. it does not aim to identify the metaphysical substance of reality, but "to make explicit already existing conceptualisations through the use of categories whose structure is influenced by natural language, the structure of human cognition and

social practices"[38]. This ontology applies the OntoClean methodology and is therefore well suited to the programme of creating a cross-disciplinary conceptualisation of information presented above. Moreover, DOLCE has been complemented by the sister ontology *Descriptions & Situations* (DnS), developed in the same original project, whose domain is the foundational modelling of different perspectives of agents on the same world events. The notion of situation is defined as an interpretation of events based on a particular conceptualisation, *i.e.* representations shared by agents and expressed by a description that assigns specific roles to the participants in the event[39].

31 Because DnS provides an ontological basis for distinguishing between events in the world and the interpretations developed by different actors of the same events, this approach has made it possible to model social roles[40], social relationships —defined as "socially-constructed objects" only interpretable within their social context[41]— and even to provide a conceptualisation of scientific communities from a constructivist point of view. This was done in a development called *Constructive Descriptions and Situations* (c.DnS). The same "facts" represented by the information can correspond to different "situations", i.e. different interpretations according to the points of view of different persons. This also applies to scientific disciplines and the different forms of knowledge they produce:

> "every scientific theory or model should be seen as a 'tool', which is the product of a specific 'knowledge collective' and whose adequacy in representing and handling specific aspects of our interaction with the world has to be tested against actual usage and effectiveness, and always be open to revision"[42].

32 The c.DnS approach, with its essential distinction between the things in the world and the objects of scientific discourse, overlaps significantly with the epistemology outlined above: "Situations are entities which represent a state of affairs, under the assumption that its components 'carve up' a view (a setting) on the domain of an ontology by virtue of a description"[43]. Furthermore, the use of notions such as "representational redescription" or "mental representations", as context-dependent views of the world by agents, situates this ontology in the broader context of the analysis of "social representations", a central concept in social psychology and sociology, as a possible foundation of social life:

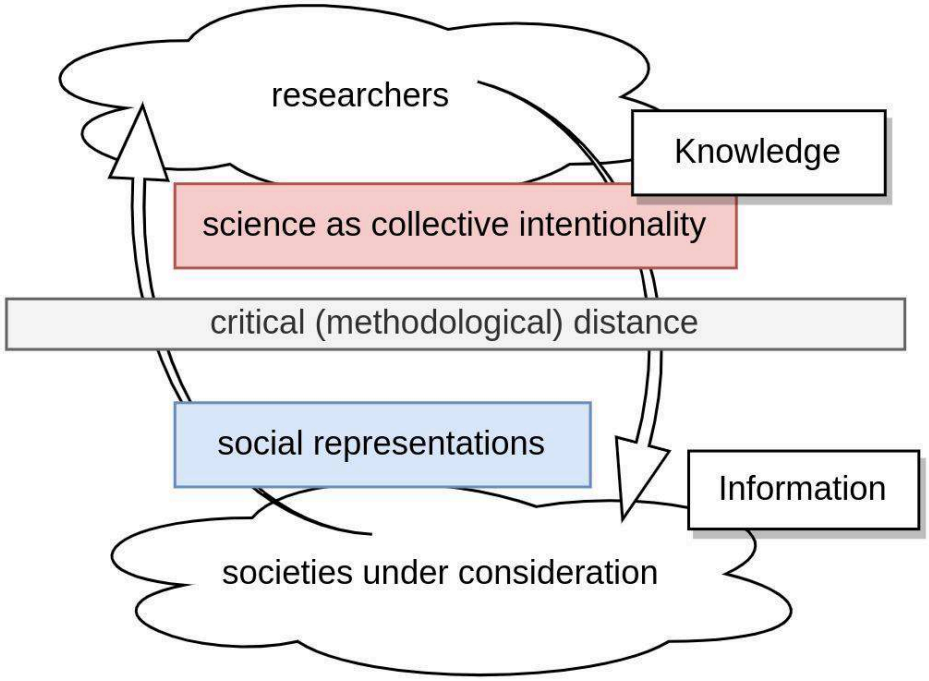
> "[Social representations] serve as the principal means for establishing and extending the shared knowledge, common practices, and affiliations that bind social members together"[44].

33 This allows c.DnS to describe "knowledge communities" as collections of "social agents that are unified by descriptions that are shared by the members". These become "intentional collectives" when they adopt a "common action schema" that unifies the collective "by means of a plan", or even they develop into "intentional normative collectives" when they share common norms. Finally, they become "knowledge collectives" insofar as they are unified by paradigms, *i.e*. "bundle-based configurations of descriptions" that express the core knowledge of scientific communities.

34 These considerations lead us to formulate two essential methodological principles in connection to our epistemological analysis (Fig. 2). The first is that scientific disciplines should be seen as structured systems of social representations: as expressions of intentional and social collectives, they are carried by scientific communities. Scientific disciplines certainly have their own rules and content, but these cannot be isolated from the social representations of the societies in which the researchers live, and evolve with them. This is particularly true of HSS, where researchers are aware of the

need for critical and methodological distance and self-reflection in order to avoid confusion and the projection of their own values onto the societies they study. This awareness also underlies the methodological approach presented above, which clearly distinguishes between two different epistemic levels: information as a representation, as factual as possible, of the society under study; knowledge as a complex model or theory that interprets the collected information about the society in question in the context of a specific discipline and the representations of a knowledge collective.

Fig. 2- Methodological distinction between social and scientific representations



The second consideration is that even if we focus on factuality, and even if different disciplines in HSS study the same things, the definition of classes in the ontology will express the conceptualisation of researchers, i.e. the collective intentionality that constitutes their discipline, and this can lead to a great diversity of views, of 'situations'. Therefore a cross-domain ontology that provides a unique and universal description of things in the world, especially about human and social life, can hardly exist. However, the same principles provide the foundation for the design of an interdisciplinary ontology based on the analysis developed by DOLCE and c.DnS. Classes are then to be understood as descriptions that represent and socially constitute scientific objects. The instances of the classes will represent things in the world, not as such but through the filter of descriptions: they are 'clipped out' of reality according to the researchers' conceptualisation because they 'satisfy the description' of the class (in the vocabulary of the c.DnS ontology). *Things* observed in the world become *scientific objects* insofar as they are identified as having the intensional properties of that class.

35 DOLCE divides particulars, i.e. the entities to which scientific discourse refers, into four distinct and non-overlapping classes: endurants, perdurants, qualities and abstracts. The essential difference between endurants and perdurants is their relationship with time: endurants, like people or artefacts, retain their identity over time, even as their

characteristics evolve; perdurants, like events or activities, develop in time and with time, and are at each moment only partially present, although identifiable as a whole. Endurants and perdurants are linked by the relation of *participation* of the former in the latter, for example, the participation of people in a meeting or a battle. We may add, and it should be noted, that perdurants, as spatio-temporal phenomena, would have no observable existence if they had no participating endurants. Furthermore, some endurants are *identifying components* of perdurants: for example, the birth of a person is not the birth of any possible person, it is the observable phenomenon of the birth of that particular person, from which it is inseparable. That specific person is an *identity-defining component* of that birth.

36 We must thus distinguish between dependent and independent objects. In the material world, this applies to a hole in a shirt, for the hole does not exist without the shirt it is in, just as a cave does not exist without the mountain it is in (these are physical features). Also the material that makes up a table (the wood, an amount of matter, or the set of planks used to build it) has an identity that is different from the one of the table itself, the latter resulting from its form, making it an identifiable physical object. In the sphere of conceptual objects, we have mental and social objects, and in particular roles and collectives, which result from the notion of classification and are analysed in the extensions of DOLCE, as we have seen. And of course, all these conceptual objects depend on persons and human collectives that create, understand and share them.

37 Two other classes, qualities and abstracts, provide a complete articulation of human discourse. *Qualities* are observable properties of endurants or perdurants that are specific to them. These include occupied space as a property of physical objects, while temporality is a property specific to events. Note that in DOLCE qualities are conceived as inherent to objects: each table has its own colour at a given moment. Each instance of the quality colour will therefore have its own value, i.e. it will occupy a point or a "region" in a reference space. *Region* is a subclass of the ontology's class *Abstracts* that comprises objects of scientific discourse which, having no temporal or spatial properties of their own, nor the status of qualities, are situated outside observable entities and, it may be added, appear to be the product of research community conventions —metric measures, for example— which allow property values to be located in a reference space.

38 If we now apply the foundational and epistemological principles that we have outlined to the analysis of information defined as a representation of factual human and social reality, the first condition to provide cross-disciplinarity in the HSS domain is the restriction of the number of *endurant* classes to a small set of disjoint classes defined by high level, rigid intensional properties. In the perspective of HSS, most objects observed in the physical world are covered by the following classes: amount of matter, inert physical natural object, individual biological object, human being, individual movable artefact, geographical place, and construction. In the intentional, social world, the most relevant endurant classes are: concept, symbol (a string, a circle), propositional content, information object (texts as collections of signs, spoken language, pictures as signs with meaning), set, human group (persons sharing a plan and virtually acting together), abstract individual (a person in fiction as a product of the human mind).

39 I will not further discuss these classes here because, as we will see, in SDHSS they are expressed using CIDOC CRM and its core extension's classes, and these can be used in information production. The point to be made here is that any further specialisation of these classes, representing characteristics of objects that are beyond those limited to strictly rigid properties used to identify them in the sense of OntoClean, should happen in the form of discipline-driven *classifications* or *extensions* of the core ontology, providing new classes and properties. Regarding classifications, discipline-specific vocabularies and taxonomies of types can be used to connotate or classify the same scientific objects, identified in their essence using the intensional properties of the foundational classes. Types can be very diverse: they can represent vocabularies of concepts used by agents in the social domain under study or adopted in the scientific classification by researchers. This distinction between specialized types that connote scientific objects, on the one side, and rigid intensional properties that define their general identity, on the other side, helps to define high-level, cross-disciplinary disjoint classes of endurants.

40 Regarding extensions, according to our methodology, the whole wealth of scientific information collected about endurants, beyond their identification, should be conceptualised in the form of qualities (colour, weight, size, etc.) or expressing their relations to other endurants, or their evolution in time and space that are observable through their participation in perdurants. Furthermore, some relations between objects only subsist in human assertions supported by common beliefs, such as saying that the author of this book is this person, and owns its copyright. The choice between representing authorship as the spatio-temporal phenomenon of the actual writing of the book manuscript, or capturing a shared belief in authorship expressed through propositional content, should be left to the different HSS disciplines and their respective research questions.

41 In other words, to properly consider the diversity of research agendas, the full richness of viewpoints from different HSS disciplines should be expressed in the domain of qualities and relationships between endurants, as well as beliefs and statements about them. It should be noted, however, that even with respect to perdurants, the presented methodology provides a set of high-level classes likely to enable cross-disciplinary interoperability. We will show how this project can be carried out by presenting the classes of the CIDOC CRM and its SDHSS extension that can be used as a cross-disciplinary ontology for the HSS domain, in line with the general view of sharing and reusing research data at the origin of this ontology[45].

Core Classes in the CIDOC CRM and the Extension Semantic Data for Humanities and Social Sciences (SDHSS)

42 The structure of the core ontology of the SDHSS ecosystem can be discovered by inspecting the tree of classes published in OntoME[46]. By progressively unfolding the tree and browsing its branches, one will find the classes and be able to access the definition of their intensional properties in the scope notes, as well as those of their relational properties establishing links to other classes. The tree shows, in addition to the CRM, the namespaces that are part of the SDHSS project. To distinguish them, I will

prefix the classes and properties with *crm* for the CRM and *sdh* for the new high-level extension.

43 The root class, *crm:E1 Entity*, contains all the objects of the CRM domain of discourse. Note that the literal values (strings, integers, geo-coordinates, etc.) are not part of it and are represented by the class *crm:E59 Primitive Value*. You therefore have to refer to existing standards, e.g. data types in the XSD namespace, for expressing them[47]. If we unfold the tree, we notice the two essential classes *crm:E77 Persistent Item* and *crm:E2 Temporal Entity*, corresponding respectively, at first glance, to the classes *Endurant* and *Perdurant* of DOLCE. However, a comparison with the DOLCE class tree shows some differences. The first thing to note is the absence of the classes *Quality* and *Abstract*, while there are four other root level classes (*crm:E54 Dimension*, *crm:E53 Place*, *crm:E52 Time Span*, *crm:E92 Spacetime Volume*). These are, in the DOLCE perspective, *regions* in time, in 2D and 3D abstract space, and in spaces of values defined by measurement units, and therefore are subclasses of *Abstract*, as they correspond to a particular position of a value in a conventional reference space. We grouped them in the extension's class *sdh:C5 Abstract Region* to avoid confusion with things existing in the observable physical world.

44 As far as the class *crm:E77 Persistent Item* and its subclasses are concerned, they include physical objects and their non-material counterparts, i.e. texts, concepts, etc. However, CRM's conceptualisation of endurants is different from DOLCE's, and some peculiarities do not conform to a strict application of OntoClean[48]. Even though the CRM has been developed by applying a precise analysis of the identity and unity of classes, the methodology that explains the taxonomies is not that of OntoClean but rather an object-oriented approach based on property analysis, understood here as the expression of relational properties between entities. As a result, the CRM has been described as a "property-centric ontology"[49]: it uses multiple inheritance in the class hierarchy and combines both those classes that are defined by *intensional* rigid properties, in the OntoClean sense, and those that provide additional qualifications in the form of *relational* properties.

45 This explains the first high-level distinction in the CRM endurants taxonomy between agents (*crm:E29 Actor*) and "inert" objects (*crm:E70 Thing*) that is based on human intentionality, with actors being persons, "individually or in groups, who have the potential to perform intentional actions". Animals and non-human agents are thus excluded from the *crm:E29 Actor* class and are modelled in the form of *crm:E24 Physical Man-Made Thing* or *crm:E20 Biological Object*, further down in the hierarchy, but there we find again, surprisingly, human beings, here understood in their biological materiality, not in their capacity of intentional action. Regarding groups (of persons) defined as actors in the CRM, they are social objects based on collective intentionality and are conceptualised in DOLCE taxonomy as belonging to the class of Agentive Social Objects, making thus a clearer difference between physical and non-physical endurants. In contrast, the distinction between essential and accidental properties of objects, notably those added in the context of human intentionality, is not applied in the CRM taxonomy of classes.

46 This is even more clear at the next stage of the endurants taxonomy, with the *crm:E72 Legal Object* class providing its descendant classes with the properties that associate these entities with the actors exercising a right over them (*crm:P105 right held by*) as well as with the right itself (*crm:P104 is subject to crm:E30 Right*), the latter being

expressed in the form of a propositional object with no explicit connection to time. We might therefore be surprised to find the *crm:E71 Human-Made Thing* class at the same level of the hierarchy as the *crm:E72 Legal Object* class, and wrongly think that this class is not a subject of law. The main reason for this modelling choice —in the property-centric approach— is to exclude some classes of conceptual objects (*crm:E55 Type*, *crm:E89:Propositional Object*) from the hierarchy of legal objects, thus not providing them with the relational properties related to rights over them. Another approach to solve this issue would be to express the notion of legal connotation in terms of an optional, time-indexed classification relation, as in DOLCE[50], which is more open to the modelling of different perspectives instead of having the *crm:E72 Legal Object* in the classes taxonomy.

47 Our analysis of the CRM identifies some restrictions, refinements and additions needed to produce a cross-disciplinary conceptualisation for the HSS research domain. Regarding endurants, only those CIDOC CRM classes defined by rigid properties in the OntoClean sense are retained in the SDHSS application profiles and used for actual data production (*crm:E24 Physical Human-Made Thing, crm:E89 Propositional Object*, etc.). Some missing classes that are relevant from the perspective of HSS research agendas, such as *sdh:C13 Geographical Place, sdh:C32 Abstract Individual, sdh:C40 Intentional Collection,* are added in the SDHSS high-level CRM extension. All other aspects of the information about these objects that are not essential to their identification should be referred to the representation of their qualities or relationships in the other main branch of the ontology, the one concerning phenomena happening in time, called *perdurants* in DOLCE and *temporal entities* in CIDOC CRM.

48 The *crm:E2 Temporal Entity* class covers all the observable phenomena that take place over a limited period of time, with an explicit reference in the scope note to the notion of *Perdurant* used in DOLCE. A careful analysis of this class shows, on the one hand, that indeed all the *crm:E2 Temporal Entity* relational properties express either a temporal relation relatively to other phenomena —in the sense of Allen's temporal properties[51]— or a relation to a *crm:E52 Time-Span* whose function is to establish a specific position of the phenomenon in the abstract time frame. On the other hand, according to DOLCE's analysis, perdurants exist by virtue of the relation of participation of endurants, which in the CRM occurs only at the level of the class *crm:E5 Event*: it is only at this level of the taxonomy that actors with the property *crm:P11 had participant* and objects with the property *crm:P12 occurred in the presence of* are explicitely associated with a temporal entity.

49 In contrast, the analysis of the *crm:E4 Period* class, which introduces in the CRM taxonomy the projection of phenomena into physical space (*crm:P8 property took place on or within* a *crm:E18 Physical Thing*) raises the issue of the definition of its essence. According to the scope note, this class covers "sets of coherent phenomena or cultural manifestations occurring in time and space", like the "Italian Renaissance". An important distinction must be made between spatio-temporal phenomena (like a birth, or the production of an artefact) that have an easily recognisable intrinsic identity, and classifications of cultural phenomena that are defined as the result of scientific research. From the perspective of the epistemological analysis outlined above, the former has the epistemic level of information, the latter of knowledge, because the identity criteria that define complex cultural phenomena result from a classification of

the observed situations that is part of an analysis that takes place in the process of scientific knowledge production.

50 From a HSS perspective, it is, therefore, preferable to consider the *crm:E5 Event* as equivalent to the DOLCE *Perdurant* class: the *crm:E5 Event* class represents spatio-temporal phenomena with an intrinsic recognisable identity and participating objects. This analysis explains the addition in the SDHSS extension of the *sdh:C3 Epistemic Situation* class, which includes "phenomena in time and physical space whose identity does not depend on intrinsic criteria but on the perspective of the observer that cuts out the situation from a more complex or long lasting phenomenon"[52]. Examples of epistemic situations are the atmospheric condition of a city during a given day, or the total economic activity of a country over the course of a year. Average temperature will be an abstracted quality of the former, while the GDP, as a quality of the total economic activity of a country during a year, will be the result of the aggregation of a number of measurements of the observed epistemic situation. The identiy of these temporal phenomena is not intrinsic but defined by the observer.

51 The *sdh:C1 Entity Quality* class is another relevant addition to the CRM. It captures the characteristics of an object at a given moment and is equivalent to DOLCE's notion of a time-indexed *Quality.* In SDHSS it is modelled as a sibling of *Perdurant* because *sdh:C1 Entity Quality* is considered to be an observable temporal phenomenon. Qualities are absent from the CRM except for *crm:E3 Condition State*. The CRM approach prefers to model qualities of objects using the *crm:E16 Measurement* class which refers to *crm:E54 Dimension* to represent a region in a quantitative abstract space defined by a unit of measurement. Note that the phenomenon captured by the class *crm:E16 Measurement* is not the quality itself but the activity of observing it, e.g. someone measuring the length of a bridge on a given day. The ontological decision to limit the information collected to the one-off observation, and to exclude from the model the intrinsic qualities of the objects, as captured by the researcher's abstraction and refinement, has the significant epistemological consequence of restricting the CRM —in this respect— to the perspective of factoids: researchers will have to inform several times in the information system the same length of this bridge that was measured at different times, whereas the aggregated factual information that such and such a bridge had this particular length during a given time-span, before its transformation and extension, is excluded from the CRM conceptualisation.

52 The *sdh:C1 Entity Quality* class thus adds an essential component to the modelling of HSS research information, allowing both the qualitative and quantitative qualities of objects, and their evolution over time, to be treated in a different and complementary way to the events that structure the CRM. The class *sdh:C1 Entity Quality,* as a subclass of *crm:E2 Temporal Entity,* has the same substance as the latter, it represents an observable phenomenon limited in time: a qualitative or quantitative property that is inseparable from the object it qualifies, with a value that characterises it at a given moment. SDHSS qualities are therefore at the same taxonomic level as the CRM events. Two properties, *sdh:P8 effects* and *sdh:P9 ends*, associate events in the spatio-temporal world with the qualities they initiate or terminate.

53 While spatio-temporal relations between objects can be expressed by instances of *crm:E5 Event* and its subclasses, purely intentional and social relations can be modelled in the CRM as instances of the *crm:E89 Propositional Object* class: the substance of the relation between the objects is then an assertion, an intentional content without a

projection into a spatio-temporal phenomenon. The crm:E30 rights class mentioned above, which only expresses propositional content, is an example of this modelling approach. But rights as collective beliefs, orders, nominations to political roles, etc., are also observable phenomena situated in time, static or dynamic, and they are central to HSS research agendas. CIDOC CRM limits its analysis of these phenomena to the events that express in materiality these intentional facts: "What goes on in our minds or is produced by our minds is also regarded as part of the material reality, as it becomes materially evident to other people at least by our utterances, behavior and products"[53]. In other words, the CRM accounts for social phenomena only by modelling their manifestation in "materiality", *i.e.* in observable spatio-temporal events. It is in this sense that the classes *crm:E66 Formation* and *crm:E68 Dissolution*, which deal with the beginning and end of the existence of groups, and *crm:E85 Joining* and *crm:E86 Leaving*, which express the relationships of actors with groups, are to be intended.

54 The CIDOC CRM thus precludes itself from modelling intentional reality as such: a class that expresses a person's membership in a group during a given period is therefore, in principle, excluded from the domain of CRM discourse. How can we then represent the political roles of people, the legal domiciles of companies, in a word, the complex properties of objects that result from social phenomena that exist only in the collective representations of people? The SDHSS extension introduces the class *sdh:C4 Intention* as a subclass of *sdh:C1 Entity Quality*, in order to integrate intentionality as envisaged by social philosophy as well as social psychology and sociology, based on the notion of mental *representations*, individual or collective as we have seen above. This notion is based on a widespread understanding in these disciplines —formulated with particular precision by the philosopher John Searle— that people, individually or in groups, pay attention to objects through their own representations[54].

55 In SDHSS, intentionality is modelled as a quality inherent to a person's mind, or, in a logic of collective intentionality, several persons', who mentally adhere to (shared) representations of objects. Intentional entities —be they humans individually or in groups, animals or even digital artefacts, according to different research agendas—can operate a classification or interpretation of an object in addition to the object's own ontological essence. This additional connotation gives the object a particular meaning, defined in the context of individual or collective representations that can be expressed as instances of the *crm:E89 Propositional Object* class [55]. Intentionality is thus conceptualised as a quality of the human brain, i.e. the presence in the brain, as a biological carrier, of propositions that belong to individual or collective representations. This individual or collective mental world underlies social life, making it possible to account for phenomena situated in time and the social space, such as the attribution of roles to persons, object ownership, group membership, etc. These phenomena are not intrinsic to the objects (persons or things) at which the intention is directed, they exist as a quality of the brains of the persons who believe the propositions about the objects to be true and thus make social reality exist.

56 The conceptualisation adopted in SDHSS is inspired by, and fits with, the ontological analysis of the *DnS* ontology around the notions of *situation* and *intentional collectives* presented above. The class *sdh:C4 Intention* thus captures the information produced by the observation of social phenomena and becomes the root of a variety of subclasses, thus acquiring a position in the SDHSS class taxonomy equivalent of that of the class *crm:E5 Event*. The coherence between the intentional level and the level of physical

materiality that grounds the CRM is established by the property *sdh:P43 has setting*, which associates the mental phenomenon with its substratum located in the sphere of spatio-temporal phenomena. Double class inheritance is also used in cases where the intentional and physical phenomena are inseparable, such as speech acts (cf. *sdh:C46 Intentional Expression*).

57 In conclusion, in view of data production and information interoperability, the SDHSS ontology ecosystem provides, on the one hand, a limited number of high-level, disjoint classes of endurants defined by rigid intensional properties. On the other hand, five main high-level classes —to be specialised in discipline-driven extensions of the ontology ecosystem— allow to organize the rich universe of research information, i.e. the characteristics and relationships of endurants that are relevant according to different research agendas: spatio-temporal events (perdurants), qualities, epistemic situations, intentionality (as temporal, individual and collective phenomenon) and its propositional content.

Conceptualizing Information Production

58 We now have all the epistemological and methodological components to describe the process of producing HSS scientific information in the form of interoperable semantic information graphs (Fig. 3). And also to be able to address the issues of identity and uncertainty that were raised in the introduction. Given the preceding analysis, this will be fairly straightforward. Our focus will not be on technical aspects, such as natural language processing or graph aggregation techniques, but on the conceptualisation of the process, in particular the application of the notions of representation and approximation.

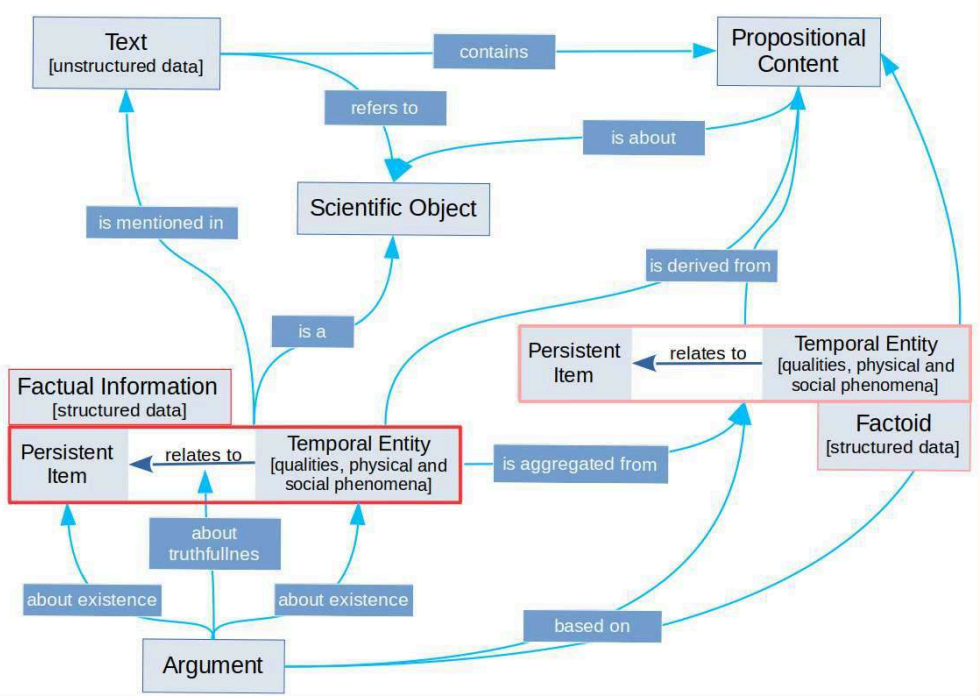
59 We will start with the case of extracting information from a text, one of the most common scenarios. Today, texts are usually available as unstructured digital data, or semi-structured data, if XML markup has been added. The first step is to identify the objects mentioned in the text. Notice that these objects are, on the one hand, things in the world and, on the other hand, scientific objects, if understood as instances of an ontology class. If we apply the foundational principles outlined above, an object will generally be an instance of a class of endurants, such as a person or an artefact, though it could also be classified as a perdurant, e.g. the set of spatio-temporal events that made up the Second World War. The same analysis can be applied to things directly observed in the world, be they artefacts in a museum, observed discussions in a self-help group, or actions carried out in the context of a craft activity, the sequence of which we want to reconstruct: we could start with direct observation of the things and then produce an information graph.

60 Let us note, first, that the identifier which denotes a thing in the world —a resource in the RDF sense— does not indicate how it is conceptualised according to an ontology. The URI provides the thing with a unique appellation in the namespace of an information system: two URIs identifying instances of different classes in different information systems can point to the same thing in the world, as the World War II example we provided in the introduction shows. We thus have to distinguish between the issue of the conceptualisation of the objects in the different information systems and the identity of the thing they represent. This is precisely the intension of the *owl:sameAs* property defined in the Web Ontology Language OWL: the "owl:sameAs

statement indicates that two URI references actually refer to the same thing: the individuals have the same 'identity' ".

61 Second, to express uncertainty about whether the string in the text really represents the intended thing, we can reify the *refers to* property using the *rdf:Statement* construct[56], adding a parameter to quantify the truthfulness of the association, in a positive or negative sense. The statement thus captures the researcher's opinion or degree of certainty about a property, which can be expressed by additional parameters about accuracy, etc., or even by a more explicit and standardized reification using the Annotation Ontology (W3C Recommendation of 23 February 2017)[57]. The RDF reification construct can also be used to express uncertainty regarding the *owl:sameAs* property, if you are unsure that two URIs point to the same thing, or for any other property in a semantic graph. The RDF-star extension of RDF[58] could also be used to identify RDF triples in order to add parameters to properties like the information source carried by the triple, its accuracy, or a truthfulness indicator defined by the researcher with the limits we'll discuss below.

Fig. 3- Process of information production



62 And third, if the URI or identifier denotes the thing, that thing becomes a scientific object when identified as an instance of a class in the ontology, in application of the epistemological principles outlined above. Therefore, adopting classes with a high level of abstraction, defined by rigid intensional properties and belonging to a common cross-disciplinary ontology, makes it easier to identify objects and to link things represented in different information systems. For example, restricting the class definition to the fact that the thing is a human-made physical artefact, and additionally providing a rich, research driven vocabulary of types, without creating more specialised subclasses, helps to identify the same object across different information systems, allowing researchers to classify it with different types: a flint object found in

an archaeological excavation could be classified as a knife or a skin scraper, but it has the same essence as a physical artefact.

63 However, this approach fails to solve all the problems. While it easily identifies people or physical artefacts, once sufficient contextual information is available, it requires more thought, and differentiation, when it comes to expressing the identity of social objects such as concepts or groups. Ultimately, it is up to the researchers to identify objects and document their choices in the information system as explicitly as possible in order to facilitate the re-use of the data.

64 Once things are identified in the text and represented as objects in the ontology, the issue becomes what qualities and relationships we want to document and how to express them in a semantic graph. This generally first happens at the epistemic level of factoids and, as we have seen, an important question arises: the semantic relations that you see in the text are those intended by the text's author or those you're interested in, even if they aren't explicitly stated (cf. above Fig. 1 and 2)? In HSS, the choice depends on the project's research agenda. Take the example of a text in a book, where an archeologist states that this particular flint artefact is a knife. If your research question is about different classifications of objects by different authors, you will want to record the archaeologist's opinion in detail. If not, you'll either ignore it or record it as an alternative to your interpretation that the artefact is a scraper.

65 The next question concerns the conceptualisation to be adopted to produce the graph and to be used to transform the propositional content of the text (as perceived in your mind) into structured data (Fig. 3). Suppose a museum record indicates that an object measures 10 cm in height. You could express this information according to the CIDOC CRM conceptualisation and add an instance of the *crm:E16 Measurement* class. The problem is that you do not really know who carried out this measurement activity, nor when. Alternatively, you could follow the SDHSS conceptualisation and create an instance of the *sdh:C44 Physical Thing Quality* class stating that the height of the object was, and probably still is 10 cm, at least during the time of existence and validity of the record. The degree of accuracy of either model depends on the information content of the source and the research question, not on the conceptualisation as such: in some situations, the former would be preferable, in others the latter.

66 As a further example, let's take a text that says that a given house belongs to a given person. This can simply be registered as a proposition, associating it with the concerned objects using relational properties in line with the CRM approach used for the *crm:E30 Right* class. Or you could create a property *owns* that associates the person's URI with the one of the house: ":person_1 :owns :house_3". But the text you're reading might provide additional information, such as the date of the text, or the period of ownership, or other details. You then have to decide whether you want to express in the graph the belief of the author of the text that the ownership subsists at that time, or conceptualise the social fact of ownership as an expression of the collective belief in the existence of the ownership at that time, as expressed by the author of the text. The last two alternatives are conceptualised in SDHSS using the *sdh:C7 Intentional State* class and its subclasses, i.e. the information that such and such propositional content —the house's ownership— is believed as valid by a person, or by the social context during a given or approximate time-span.

67 With regard to information formalisation, this example highlights three different approaches commonly used in the Semantic Web. The first is based on simple relational

properties, in this case, the *owns* property. This approach can express the fact of ownership but does not capture all the available information. The second is the reification in the form of statements, which is at the heart of the Wikidata information system[59]. In this community driven knowledge graph you don't fill in properties in the RDF sense, but statements (similar to *rdf:Statement* but stored in a relational database) identified by a URI. This allows additional information to be added to the relational property, such as the date of the fact or its source. We could adopt the RDF-star construct with the same logic, adding dates to the triple in order to express the time-span of the ownership (that such and such a person owns such and such a house at that time according to that document). However, this approach is not recommended by the authors of the RDF-star specification because it is designed to provide metadata about triples and not to provide a complex model of information. If a triple represents different situations, at different time-spans, given that the identity of the triple is always the same in RDF, you would not be able to distinguish between different situations in the world because they would be expressed by the same triple[60].

68 Consider the example of the property *work location* used in Wikidata to represent different locations where Johannes Kepler was active[61]. The relation between the astronomer and the geographical place is enriched not only with the metadata about the source but also with the concerned time-spans. This additional information is stored as statements about binary relationships. But how would you express the fact that Kepler worked twice in the same place, but at different times, or in different institutions at different times but in the same geographical place? You could not represent these different facts using embedded triples in RDF-Star, because they would be expressed with different predicates about the same, identical triple, with no way of distinguishing the different facts represented, since the identity of a triple consists in its three terms, and they would be the same. In Wikidata, or the *rdf:Statement* construct, or other forms of reification, this could work because you would provide different identifiers to the same statements/triples[62]. However, the real problem here is that we are trying to use statements, *i.e.* binary predicates, to express situations that are *de facto* n-ary and involve multiple objects in the same temporal situation.

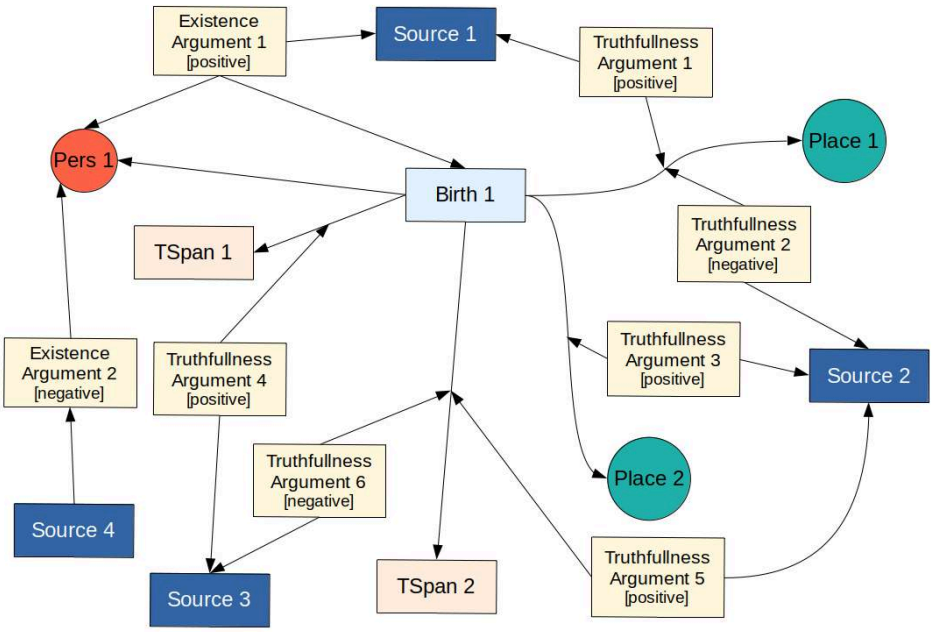
69 Furthermore, to continue the same example, researchers could consider it more relevant to first know in which institution Kepler worked and then that this activity happened at that place. If we only use binary relationships, in one information system we could give priority, for the same situation, to the relationship between the person and the place of activity, while adding the institution concerned as an accessory, whereas in another information system, the main relationship, expressed by the another statement, would be that between the person and the institution, with the place being relegated to the status of a complement, even though in both cases the situation is exactly the same. This is not the best way to promote data interoperability across different information systems.

70 From the point of view of producing FAIR research data that facilitate their reuse for new research agendas, it seems preferable to choose the third approach present in the Semantic Web, that of a consequent reification of the relationships between objects, and their qualities: they will be conceptualised as temporal phenomena, be they physical or intentional, or as epistemic situations as we have outlined in our foundational analysis based on DOLCE and DnS. In the former example, we will have to create multiple instances of the *crm:E7 Activity* class, or a more specific subclass, to

represent the different segments of Kepler's activity. Each instance of this spatio-temporal phenomenon will have its own identifier, and all the related objects (institutions and geographical places) will be associated by distinct relational properties.

71 This approach has a number of benefits. It allows the development of a richer, more explicit conceptualisation, in relation to different research agendas and points of view, which makes research data easier to reuse. We can formalise more easily, through the associated relational properties, the role of the different objects involved in the information. In the case of uncertainty due to conflicting statements in the sources, we can provide alternative relations to different objects for the same property and, using statements, provide the source and degree of truthfulness for each alternative. Figure 4 provides an example: you could provide two alternative birthplaces for a person's birth and add arguments about the alternative statements' different truthfulness values (positive or negative) regarding the different sources. This method, endorsed by the CIDOC CRM[63], can also be applied to dates, although the CRM provides a robust conceptualisation for dealing with time uncertainty that is directly integrated into the ontology: on the one hand, with properties that provide inner and outer boundaries of temporal phenomena, in order to take into account not only uncertainty but also the general fuzziness of the boundaries of temporal phenomena; on the other hand, with properties that express the relative position of events to each other even if the precise location on the time scale is not known[64].

Fig. 4- Arguments and alternatives



It should be noted, however, that according to the foundational analysis presented above, there are statements for which an alternative between associated objects is not possible because the objects in question constitute the identity of the temporal phenomenon. This applies to qualities inherent in the object they qualify but also to spatio-temporal phenomena such as a birth. The person associated with the birth admits of no alternative because it is his or her own birth. An instance of the *crm:E67*

Birth class is not the birth of just anyone, but the representation of the spatio-temporal event in which this specific person participated (Fig. 4). The objects associated with temporal phenomena using this kind of properties are considered in SDHSS as identity defining for the concerned phenomena[65]. Finally, arguments can also qualify objects, and not just statements, but in this case, we are not discussing truthfulness, which is a property of statements, but the objects' existence: did this person really exist or is it just a fictional character? Did this person really have this social role or was it just pretended in order to get a higher position? Existence arguments about things and facts in the world make it possible to express considerations regarding their plausibility, in relation to the existing documentation and the confidence that can be placed in it.

72 When we discuss arguments about existence and truthfulness, and statements about alternative objects that may have the same property, we implicitly are at the epistemic level of factual information. About factoids you cannot have alternative statements because they represent in form of structured data propositions identified in a document by a researcher. For each property associating an object to a factoid, you would just have one assertion in the text, e.g. just one place of birth. As we have seen, moving from the epistemic level of factoids to that of factual information requires a process of aggregation and merging of redundant and sometimes contradictory information.

73 This integration process can be accomplished according to two different scenarios. In the first one, the identity defining object for a temporal phenomenon —in our example the person for his or her birth— is well-known and identifiable from the context: this helps to merge the different birth factoids into one factual birth and provide multiple statements if some related objects, like the place of birth, differ. We could also provide truthfulness values according to the quality of the source and the reliability of the information extraction process. Note that this aggregation process concerns not only factoids extracted from texts but also sources of structured data available on the Semantic Web. From the researcher's point of view, these sources, however good they may be, all have the status of factoids, particularly because of the possible redundancy of the information. As we have seen, the researcher's information system must ultimately provide a consistent factual representation of the world under study, to enable the answer of research questions. This underlines the importance of LODs and authority files. Linking objects that have been identified with the highest possible degree of accuracy greatly facilitates the process of aggregation of the disparate sources of information available on the Semantic Web to produce high-quality, rich factual information.

74 In the second scenario, the objects involved in the temporal phenomena are not yet identified: for example, you can retrieve the mention of someone's birth but you don't yet know who this person is. In this case, in order to identify the participating things, you will have to define patterns that allow you to produce factoids as temporal classes and associate to them already identified objects (e.g. concepts, institutions, geographical places), in association with dates and similarity of labels denoting the object in order to identify it: if the things are the same, their qualities and relationships, as well as dates and appellations, must have a number of similarities, according to a given threshold. Also, because the information is available on the Semantic Web according to the three aforementioned different formalisations —

relational properties, statements and temporal phenomena—, it will be necessary to choose between two standardisation strategies: either an overhaul of all the information in the form of binary relations, with the advantage of simplicity but the disadvantage of impoverishing the information; or a transformation of all the data towards a temporal-phenomena-centred model, albeit with gaps, but allowing to mobilise all the information available for record linkage. Discussion of the most effective strategy is outside the scope of this paper.

75 These two strategies will make it possible to identify a certain number of objects: the provisional identifiers assigned previously to not yet identified objects will then be replaced with the URI retained in the main record for an object, both in the annotated texts and the information graph. Once this step has been completed, we can proceed with the process of aggregating the factoids into factual, non-redundant information for all available qualities and relationships according to the first scenario mentioned above. In doing so, we can express the source of each piece of information, and the accuracy of the information extracted, in the statements' metadata and/or through existence and truthfulness arguments. Indeed, we must not forget that scientific information, in the form of digital data, is not only a representation but also an approximation of factual reality. For this reason, it is essential to document its quality using the PROV-O ontology[66] and its Hic-O[67] extension, or any other standard, allowing us to share metadata about the information origin and quality. The result will be an information graph of hitherto unimaginable volume, whose quality and robust, extensible conceptualisation will be clearly documented, making it easier to reuse data for new research in line with FAIR principles.

Conclusion

76 In the introduction to this paper, we set out three challenges to the vision of a distributed, cumulative, giant semantic graph, collaboratively maintained by HSS researchers and reusable for new research: the difficulty of proposing a cross-disciplinary ontology for HSS given the different research agendas and constructivist approaches; the difficulty of identifying and interlinking the objects present in the different information systems; the challenge of creating and sharing 'factual' information, when the sources of information are uncertain, inconsistent, and contradictory.

77 An epistemological analysis of information production in HSS shows that factual information is at the core of data interoperability and reuse for new research. Factual information must be understood as the best possible representation of objects in the world, their properties and relationships. Although different disciplines model the same things from different perspectives, we showed that a foundational analysis of the objects of scientific discourse using the DOLCE and DnS ontologies allows us to define high-level, disjoint classes of objects that can ground interoperability, and this not only in the domain of *endurants* (persons, groups, artefacts, concepts, etc.), but also in the domain of temporal phenomena (*perdurants*), be they qualities of objects, or their relationships in the physical or social space, or epistemic situations. We propose to use the CIDOC CRM and its SDHSS integration as a cross-disciplinary core ontology for the HSS and to add extensions related to specific disciplinary domains, based on the outlined foundational analysis.

78 This approach helps to facilitate the identification of objects present in different semantic graphs: if they are produced as instances of a limited number of classes defined by rigid properties, in the sense of OntoClean, they will be more easily interlinked, letting the different disciplines and projects to classify these objects with more specialised and rich, domain related controlled vocabularies. Regarding the issue of identity, the distinction between factoids —as representations of the content of sources in the form of structured data— and factual information, as aggregated factoids, allows us to ground the methodology for a progressive identification of objects during the process of factual information production.

79 Finally, we have shown that the CIDOC CRM integrates methods for expressing uncertainty, and in particular temporal uncertainty, with different properties to express the fuzziness of the boundaries of temporal phenomena, or their relative position in the flow of time when a precise date is not known. In case of contradictions in documents or existing semantic graphs, the adoption of a temporal phenomena-centred data structure allows us to add alternative statements about the same properties and related objects, thus capturing the full complexity of research information. Arguments about the existence of an object, or its fictional nature, or the truthfulness of a statement, make it possible to document the accuracy of the available information and its degree of approximation of factuality.

80 In conclusion, these challenges do not appear to be an insurmountable obstacle to the realisation of the "giant knowledge graph" of HSS. On the contrary, the epistemological, foundational and methodological principles we have outlined enable us to embark on two major undertakings. On the one hand, the implementation of collaborative information systems aimed at facilitating access to these information production methods and improving the quality of information through appropriate virtual research environments. That's why we decided to build the Geovistory platform, and we hope that similar initiatives will emerge for other disciplines or contexts. On the other hand, we need to integrate recent developments in AI, especially in the fields of machine learning, neural networks and LLM, in order to automate and facilitate as much as possible the process of verifying and aggregating information, as well as identifying the same objects in different corpora and semantic graphs. These are the urgent challenges that need to be addressed in order to realise the project of a distributed, sustainable, giant information graph for HSS research: overcoming these challenges will require establishing a culture of sharing data, skills and resources across disciplines, thereby contributing to a critical digital transition in the HSS.

BIBLIOGRAPHY

Agazzi, Evandro (2014), *Scientific Objectivity and Its Contexts*, Cham, Springer, https://doi.org/10.1007/978-3-319-04660-0.

Agazzi, Evandro (2017), "The Truth of Theories and Scientific Realism", in Evandro Agazzi (ed.), *Varieties of Scientific Realism: Objectivity and Truth in Science*, Cham, Springer.

Akoka, Jacky, Isabelle Comyn-Wattiau *et al*. (2020), "Contribution of Conceptual Modeling to Enhancing Historians' Intuition - Application to Prosopography", in Dobbie Gillian, Frank Ulrich, *et al*. (eds.), *Conceptual Modeling*, Cham, Springer International Publishing, p. 164-173, https://doi.org/10.1007/978-3-030-62522-1_12.

Beretta, Francesco (2021), "A Challenge for Historical Research: Making Data FAIR Using a Collaborative Ontology Management Environment (OntoME)", *Semantic Web* 12.2, p. 279-294, https://doi.org/10.3233/SW-200416.

Beretta, Francesco (2022), « Interopérabilité des données de la recherche et ontologies fondationnelles : un écosystème d'extensions du CIDOC CRM pour les sciences humaines et sociales », in Nicolas Lasolle, Olivier Bruneau & Jean Lieber (éds.), *Actes des journées Humanités Numériques et Web sémantique*, Nancy, France, p. 2-22, https://doi.org/10.5281/zenodo.7014341.

Beretta, Francesco (2023), « Données ouvertes liées et recherche historique : un changement de paradigme », *Humanités numériques* 7, https://doi.org/10.4000/revuehn.3349.

Beretta, Francesco (2024). "An ontology of geographical places and their spatio-temporal, social evolution in the context of an ecosystem of CIDOC CRM extensions for humanities and social sciences (SDHSS)", in Wiesława Duży (ed.), *Modelling the City. Formal Ontology and Spatial Humanities*, Routledge, p. 21-45, https://doi.org/10.4324/9781032695891.

Beretta, Francesco (2024), « Données liées ouvertes et référentiels public : un changement de paradigme pour la recherche en sciences humaines et sociales », *Arabesques* 112, p. 26-27, https://doi.org/10.35562/arabesques.3820.

Beretta, Francesco (2024), "Semantic Data for Humanities and Social Sciences (SDHSS): An Ecosystem of CIDOC CRM Extensions for Research Data Production and Reuse", in Thomas Riechert, Hartmut Beyer, Jennifer Blanke & Edgard Marx (eds.), *Professorale Karrieremuster*, Leipzig, Open-Access-Hochschulverlag, p. 73-102. https://doi.org/10.33968/9783966270502-05.

Borgo, Stefano & Claudio Masolo (2009), "Foundational Choices in DOLCE", in Steffen Staab & Rudi Studer (eds.), *Handbook on Ontologies*, 2nd ed., Berlin, Heidelberg, Springer, p. 361-381.

Borgo, Stefano & al. (2022), "Foundational Ontologies in Action", *Applied Ontology* 17 (1), https://doi.org/10.3233/AO-220265.

Borgo, Stefano, *et al.* (2022), "DOLCE: A Descriptive Ontology for Linguistic and Cognitive Engineering", *Applied Ontology* 17(1), https://doi.org/10.3233/AO-210259.

Champin, Pierre-Antoine (2022), "RDF-star patterns for provenance", 26 January 2022, https://www.w3.org/community/rdf-dev/2022/01/26/provenance-in-rdf-star/.

Daquino, Marilena & Francesca Tomasi (2015), "Historical Context Ontology (HiCO): A Conceptual Model for Describing Context Information of Cultural Heritage Objects", in Emmanouel Garoufallou, Richard J. Hartley & Panorea Gaitanou (eds.), *Metadata and Semantics Research*, Cham, Springer International Publishing, p. 424-436, https://doi.org/10.1007/978-3-319-24129-6_37.

Doerr Martin (2003), "The CIDOC Conceptual Reference Module: An Ontological Approach to Semantic Interoperability of Metadata", *AI Magazine* 24(3), p. 75-92, https://doi.org/10.1609/aimag.v24i3.1720.

Doerr Martin & Dolores Iorizzo (2008), "The dream of a global knowledge network - A new approach", *Journal on Computing and Cultural Heritage* 1 (1), p. 1-23, https://doi.org/10.1145/1367080.1367085.

Doerr, Martin *et al.* (eds) (2022), *Definition of the CIDOC Conceptual Reference Model*, https://cidoc-crm.org/Version/version-7.1.2.

Edmond, Jennifer (2019), "Strategies and Recommendations for the Management of Uncertainty in Research Tools and Environments for Digital History", *Informatics* 6 (3), https://doi.org/10.3390/informatics6030036.

Gangemi, Aldo (2008), "Norms and plans as unification criteria for social collectives", *Autonomous Agents and Multi-Agent Systems* 17.1 2008, p. 70-112, https://doi.org/10.1007/s10458-008-9038-9.

Gangemi, Aldo & Peter Mika (2003), "Understanding the Semantic Web through Descriptions and Situations", *On The Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE*, vol. 2888, Berlin, Heidelberg, Springer Berlin Heidelberg, p. 689-706, https://doi.org/10.1007/978-3-540-39964-3_44.

Gangemi Aldo & Valentina Presutti (2016), "Dolce+D&S Ultralite and Its Main Ontology Design Patterns", in Pascal Hitzler *et al.* (eds), *Ontology Engineering with Ontology Design Patterns: Foundations and Applications,* Amsterdam Berlin, IOS Press AKA, p. 81-103.

Garbacz Pawel, Bogumił Szady & Agnieszka Ławrynowicz (2021), "Identity of Historical Localities in Information Systems", *Applied Ontology* 16.1, p. 55-86, https://doi.org/10.3233/AO-200235.

Gruber, Tom (2018), "Ontology", in Ling Liu & M. Tamer Özsu (eds.), *Encyclopedia of Database Systems*, 2nd edition, New York, Springer, p. 2574-2576, https://doi.org/10.1007/978-1-4614-8265-9.

Guarino, Nicola & Christopher A. Welty (2009), "An Overview of OntoClean", in Steffen Staab & Rudi Studer (eds.), *Handbook on Ontologies*, 2nd ed., Berlin, Heidelberg, Springer.

Guizzardi, Giancarlo (2020), "Ontology, Ontologies and the 'I' of FAIR", *Data Intelligence*, 2 (1–2), p. 181-191, https://doi.org/10.1162/dint_a_00040.

Giancarlo Guizzardi *et al.* (2022), "UFO: Unified Foundational Ontology", *Applied Ontology*, 17.1, https://doi.org/10.3233/AO-210256.

Guizzardi Giancarlo & Nicola Guarino (2023), "Semantics, Ontology and Explanation", https://doi.org/10.48550/arXiv.2304.11124.

Hitzler, Pascal (2010), Markus Krötzsch & Sebastian Rudolph, *Foundations of Semantic Web Technologies*, Boca Raton, CRC Press.

Holmen Jon & Christian-Emil Ore (2010), "Deducing Event Chronology in a Cultural Heritage Documentation System", in Bernard Frischer *et al.* (eds.) *Making History Interactive. Computer Applications and Quantitative Methods in Archaeology (CAA),* Oxford, Archaeopress, p. 122-129.

Jacob, Pierre (2023), "Intentionality", in Zalta Edward N. & Uri Nodelman (eds.), *The Stanford Encyclopedia of Philosophy*, Spring 2023, Metaphysics Research Lab, Stanford University, https://plato.stanford.edu/archives/spr2023/entries/intentionality/.

Levine-Clark, Michael & John D. McDonald (eds) (2019), *Encyclopedia of Library and Information Sciences*, 4th ed., Boca Raton, CRC Press, https://doi.org/10.1081/E-ELIS4.

Masolo, Claudio *et al.* (2003), "Descriptive Ontology for Linguistic and Cognitive Engineering (DOLCE)", *WonderWeb Deliverable D18. Ontology Library,* Laboratory for Applied Ontology, Trento.

Masolo, Claudio *et al.* (2004), "Social Roles and their Descriptions", in *Principles of Knowledge Representation and Reasoning*, Whistler, Canada, p. 267-277, http://www.aaai.org/Library/KR/2004/kr04-029.php.

McKinney, Earl H. & Charles J. Yoos (2010), "Information about Information: A Taxonomy of Views", *MIS Quarterly* 34 (2), p. 329-344, https://doi.org/10.2307/20721430.

Mika Peter & Aldo Gangemi (2004), "Descriptions of social relations", in *Proceedings of the 1st Workshop on Friend of a Friend, Social Networking and the Semantic Web*, Galway, https://www.w3.org/2001/sw/Europe/events/foaf-galway/papers/fp/descriptions_of_social_relations/.

Orlandi Fabrizio, Damien Graux & Declan O'Sullivan (2021), "Benchmarking RDF Metadata Representations: Reification, Singleton Property and RDF", in *2021 IEEE 15th International Conference on Semantic Computing (ICSC)*, p. 233-240, https://doi.org/10.1109/ICSC50631.2021.00049.

Pasin, Michele & John Bradley (2015), "Factoid-Based Prosopography and Computer Ontologies: Towards an Integrated Approach", *Literary and Linguistic Computing*, 30.1, p. 86-97, https://doi.org/10.1093/llc/fqt037.

Rowley, Jennifer E. (2007), "The Wisdom Hierarchy: Representations of the DIKW Hierarchy", *Journal of Information Science* 33.2, p. 163–80, https://doi.org/10.1177/0165551506070706.

Sammut, Gordon & Caroline Horwath (2014), "Social Representations", in Thomas Teo (ed.), *Encyclopedia of critical psychology,* New York, Springer Reference.

Sammut, Gordon *et al.* (eds.) (2015), *The Cambridge Handbook of Social Representations,* Cambridge, University Press.

Sanfilippo, Emilio M. & Beatrice Markhoff (2020), "Ontological Analysis and Modularization of CIDOC-CRM", *Formal Ontology in Information Systems*, p. 107-121, https://doi.org/10.3233/FAIA200664.

Schweikard David P. & Hans Bernhard Schmid (2021), "Collective Intentionality", in Zalta Edward N. (éd.), *The Stanford Encyclopedia of Philosophy*, Fall 2021, Metaphysics Research Lab, Stanford University, https://plato.stanford.edu/archives/fall2021/entries/collective-intentionality/.

Schneider, Philipp, Jim Jones, Torsten Hiltmann & Tomi Kauppinen (2021), "Challenge-Derived Design Practices for a Semantic Gazetteer for Medieval and Early Modern Places", *Semantic Web*, 12.3, p. 493-515 https://doi.org/10.3233/SW-200394.

Searle, John (2010), *Making the Social World: The Structure of Human Civilization,* Oxford, University Press, https://doi.org/10.1093/acprof:osobl/9780195396171.001.0001.

Sikos, Leslie F. & Philp Dean (2020), "Provenance-Aware Knowledge Representation: A Survey of Data Models and Contextualized Knowledge Graphs", *Data Science and Engineering* 5(3), p. 293-316, https://doi.org/10.1007/s41019-020-00118-0.

Trojahn Cassia et al. (2022), "Foundational ontologies meet ontology matching: A survey", *Semantic Web* 13 (4), p. 685-704. https://doi.org/10.3233/SW-210447.

Van Campenhoudt, Luc, Jacques Marquet & Raymond Quivy (2022), *Manuel de recherche en sciences sociales*, 6e édition, Paris, Armand Colin.

NOTES

1. In addition to the corresponding entries in Wikipedia, see the "Linked Data", "Semantic Interoperability" and "Semantic Web" entries in Michael Levine-Clark & John D. McDonald (eds) (2019), *Encyclopedia of Library and Information Sciences*, 4th ed., Boca Raton, CRC Press.

2. https://en.wikipedia.org/wiki/Google_Knowledge_Graph. This sentence, available in September 2023, has been removed from Wikipedia and only the previous values, accompanied by a source, can be found there: "By May 2020, this had grown to 500 billion facts on 5 billion entities" (December 2023).

3. Francesco Beretta (2022), « Interopérabilité des données de la recherche et ontologies fondationnelles : un écosystème d'extensions du CIDOC CRM pour les sciences humaines et sociales », in Nicolas Lasolle, Olivier Bruneau & Jean Lieber (éds.), *Actes des journées Humanités Numériques et Web sémantique*, Nancy, France, p. 2-22. ; Francesco Beretta (2024), "Semantic Data for Humanities and Social Sciences (SDHSS): An Ecosystem of CIDOC CRM Extensions for Research Data Production and Reuse", in Thomas Riechert, Hartmut Beyer, Jennifer Blanke & Edgard Marx (eds.), *Professorale Karrieremuster*, Leipzig, Open-Access-Hochschulverlag, p. 73-102.

4. Tom Gruber (2018), "Ontology", in Ling Liu & M. Tamer Özsu (eds.), *Encyclopedia of Database Systems*, 2nd edition, New York, Springer, p. 2574-2576.

5. https://ontome.net.

6. Francesco Beretta (2021), "A Challenge for Historical Research: Making Data FAIR Using a Collaborative Ontology Management Environment (OntoME)", *Semantic Web* 12.2, p. 279-294.

7. Tom Gruber (2018), "Ontology".

8. Pascal Hitzler, Markus Krötzsch & Sebastian Rudolph (2010), *Foundations of Semantic Web Technologies*, Boca Raton, CRC Press, p. 66-67.

9. https://www.geovistory.org/. Further details below.

10. Francesco Beretta (2024), « Données liées ouvertes et référentiels public : un changement de paradigme pour la recherche en sciences humaines et sociales », *Arabesques* 112, p. 26-27.

11. https://ontome.net/namespace/11.

12. Francesco Beretta (2023), « Données ouvertes liées et recherche historique : un changement de paradigme », *Humanités numériques* 7. http://dx.doi.org/10.4000/revuehn.3349

13. Ron Mallon (2019), "Naturalistic Approaches to Social Construction", in Edward N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy,* https://plato.stanford.edu/archives/spr2019/entries/social-construction-naturalistic/.

14. https://www.wikidata.org/wiki/Q362.

15. https://www.wikidata.org/wiki/Q1190554.

16. https://catalogue.bnf.fr/ark:/12148/cb11996115g.

17. https://www.w3.org/TR/owl-ref/#sameAs-def.

18. https://www.wikidata.org/wiki/Q122962941.

19. https://www.google.com/search?kgmid=/g/11l36k_n9f.

20. Cf. Jennifer Edmond (2019), "Strategies and Recommendations for the Management of Uncertainty in Research Tools and Environments for Digital History", *Informatics* 6 (3).

21. For a discussion of these issues in relation to the identification of geographical places, cf. Pawel Garbacz, Bogumił Szady & Agnieszka Ławrynowicz (2021), "Identity of Historical Localities in Information Systems", *Applied Ontology* 16.1, p. 55-86; Philipp Schneider, Jim Jones, Torsten Hiltmann & Tomi Kauppinen (2021), "Challenge-Derived Design Practices for a Semantic Gazetteer for Medieval and Early Modern Places", *Semantic Web*, 12.3, p. 493-515.

22. https://www.geovistory.org/.

23. Jennifer E. Rowley (2007), "The Wisdom Hierarchy: Representations of the DIKW Hierarchy", *Journal of Information Science* 33.2, p. 163-180, https://doi.org/10.1177/0165551506070706.

24. Luc Van Campenhoudt, Jacques Marquet & Raymond Quivy (2022), *Manuel de recherche en sciences sociales*, 6ème édition, Paris, Armand Colin.

25. Earl H. McKinney & Charles J. Yoos (2010), "Information About Information: A Taxonomy of Views", *MIS Quarterly* 34 (2), p. 329-344.

26. Michele Pasin & John Bradley (2015), "Factoid-Based Prosopography and Computer Ontologies: Towards an Integrated Approach", *Literary and Linguistic Computing*, 30 (1), p. 86-97.

27. For a model of the scientific process in historical research, from factoids production to hypothesis testing, see Jacky Akoka, Isabelle Comyn-Wattiau *et al.* (2020), "Contribution of Conceptual Modeling to Enhancing Historians' Intuition - Application to Prosopography", in

Dobbie Gillian, Frank Ulrich, *et al.* (eds.), *Conceptual Modeling*, Cham, Springer International Publishing, p. 164-173,

28. https://en.wikipedia.org/wiki/Falsifiability.

29. Tom Gruber (2018), "Ontology".

30. Nicola Guarino & Christopher A. Welty (2009), "An Overview of OntoClean", in Steffen Staab & Rudi Studer (eds.), *Handbook on Ontologies,* 2nd ed., Berlin, Heidelberg, Springer.

31. Francesco Beretta (2024), "An ontology of geographical places and their spatio-temporal, social evolution in the context of an ecosystem of CIDOC CRM extensions for humanities and social sciences (SDHSS)", in Wiesława Duży (ed.), *Modelling the City. Formal Ontology and Spatial Humanities*, Routledge. p. 21-45.

32. Evandro Agazzi (2017), "The Truth of Theories and Scientific Realism", in Evandro Agazzi (ed.), *Varieties of Scientific Realism: Objectivity and Truth in Science*, Cham, Springer, p. 49–68. Cf. Evandro Agazzi (2014), *Scientific Objectivity and Its Contexts*, Cham, Springer, https://doi.org/10.1007/978-3-319-04660-0.

33. Tom Gruber (2018), 'Ontology'.

34. Giancarlo, Guizzardi & Nicola Guarino (2023), "Semantics, Ontology and Explanation", https://doi.org/10.48550/arXiv.2304.11124.

35. Giancarlo Guizzardi (2020), "Ontology, Ontologies and the 'I' of FAIR", *Data Intelligence*, 2 (1–2), p. 181-191.

36. Stefano Borgo *et al.* (2022), "Foundational Ontologies in Action", *Applied Ontology* 17 (1); Trojahn Cassia *et al.*, (2022), "Foundational ontologies meet ontology matching: A survey", *Semantic Web* 13 (4), p. 685-704.

37. Giancarlo Guizzardi *et al.* (2022), "UFO: Unified Foundational Ontology", *Applied Ontology*, 17.1.

38. Claudio Masolo *et al.* (2003), "Descriptive Ontology for Linguistic and Cognitive Engineering (DOLCE)", *WonderWeb Deliverable D18. Ontology Library,* Laboratory for Applied Ontology, Trento. Stefano Borgo & Claudio Masolo (2009), "Foundational Choices in DOLCE", in Steffen Staab & Rudi Studer (eds.), *Handbook on Ontologies*, 2nd ed., Berlin, Heidelberg, Springer, p. 361-381.

39. Claudio Masolo *et al.*, "An Ontology of Descriptions and Situations", *WonderWeb deliverable D18. Ontology library,* Laboratory for Applied Ontology, Trento p. 97-99. Aldo Gangemi & Mika Peter (2003), "Understanding the Semantic Web through Descriptions and Situations", *On The Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE*, vol. 2888, Berlin, Heidelberg, Springer Berlin Heidelberg, p. 689-706. Aldo Gangemi & Valentina Presutti (2016), "Dolce+D&S Ultralite and Its Main Ontology Design Patterns", in Pascal Hitzler et al. (eds), *Ontology Engineering with Ontology Design Patterns: Foundations and Applications,* Amsterdam Berlin, IOS Press AKA, p. 81-103.

40. Claudio Masolo *et al.* (2004), "Social Roles and their Descriptions", in *Principles of Knowledge Representation and Reasoning*, Whistler, Canada, p. 267-277.

41. Peter Mika & Aldo Gangemi (2004), "Descriptions of social relations", in *Proceedings of the 1st Workshop on Friend of a Friend, Social Networking and the Semantic Web, Galway*.

42. Aldo Gangemi (2008), "Norms and plans as unification criteria for social collectives", *Autonomous Agents and Multi-Agent Systems* 17.1, p. 70-112.

43. Aldo Gangemi (2008), "Norms and plans".

44. Gordon Sammut & Caroline Horwath (2014), "Social Representations", in Thomas Teo (ed.), *Encyclopedia of critical psychology,* New York, Springer Reference (see also entries: Interobjectivity; Social Constructionism; Social Representations; Socialization); Gordon Sammut *et al.* (eds.) (2015), *The Cambridge Handbook of Social Representations,* Cambridge, University Press.

45. Martin Doerr & Dolores Iorizzo (2008), "The dream of a global knowledge network—A new approach", *Journal on Computing and Cultural Heritage* 1 (1), p. 1-23.

46. https://ontome.net/classes-tree. The tree can be browsed without a login.

47. https://www.w3.org/TR/xmlschema-2/.

48. Emilio M. Sanfilippo & Beatrice Markhoff (2020), "Ontological Analysis and Modularization of CIDOC-CRM", *Formal Ontology in Information Systems*, p. 107-121.
49. Martin Doerr (2003), "The CIDOC Conceptual Reference Module: An Ontological Approach to Semantic Interoperability of Metadata", *AI Magazine* 24 (3), p. 75-92.
50. Cf. the "classification principle" in Stefano Borgo *et al.* (2022), "DOLCE: A Descriptive Ontology for Linguistic and Cognitive Engineering", *Applied Ontology*, 17(1), https://doi.org/10.3233/AO-210259.
51. Jon Holmen & Christian-Emil Ore (2010), "Deducing Event Chronology in a Cultural Heritage Documentation System", in Bernard Frischer *et al.* (eds.) *Making History Interactive. Computer Applications and Quantitative Methods in Archaeology (CAA),* Oxford, Archaeopress, p. 122-129.
52. If there are no other references, the scope notes of the respective classes are quoted.
53. Martin Doerr *et al.* (eds) (2022), *Definition of the CIDOC Conceptual Reference Model*, p. 25, https://cidoc-crm.org/Version/version-7.1.2.
54. John Searle (2010), *Making the Social World: The Structure of Human Civilization*, Oxford, University Press.
55. Pierre Jacob (2023), "Intentionality", in Edward N Zalta. & Uri Nodelman (eds.), *The Stanford Encyclopedia of Philosophy*, Spring 2023, Metaphysics Research Lab, Stanford University, https://plato.stanford.edu/archives/spr2023/entries/intentionality/; David P. Schweikard & Hans Bernhard Schmid (2021), "Collective Intentionality", in Edward N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy*, Fall 2021, Metaphysics Research Lab, Stanford University, https://plato.stanford.edu/archives/fall2021/entries/collective-intentionality/.
56. https://www.w3.org/TR/2004/REC-rdf-primer-20040210/#reification.
57. https://www.w3.org/TR/annotation-vocab/#annotation.
58. https://w3c.github.io/rdf-star/cg-spec/2021-12-17.html.
59. https://www.wikidata.org/wiki/Help:Statements.
60. Pierre-Antoine Champin (2022), "RDF-star patterns for provenance", 26 January 2022, https://www.w3.org/community/rdf-dev/2022/01/26/provenance-in-rdf-star/.
61. https://www.wikidata.org/wiki/Q8963.
62. Fabrizio Orlandi, Damien Graux & Declan O'Sullivan (2021), "Benchmarking RDF Metadata Representations: Reification, Singleton Property and RDF", in *2021 IEEE 15th International Conference on Semantic Computing (ICSC)*, p. 233-240.
63. Martin Doerr *et al.* (eds) (2022), *Definition of the CIDOC Conceptual Reference Model*, p. 23.
64. Martin Doerr *et al.* (eds) (2022), *Definition of the CIDOC Conceptual Reference Model*, p. 43-6.
65. This is made explicit in the SDHSS ecosystem by making them subproperties of the *sdh:P2 domain class has identity defining component* property.
66. https://www.w3.org/TR/prov-o/; Leslie F. Sikos & Philp Dean (2020), "Provenance-Aware Knowledge Representation: A Survey of Data Models and Contextualized Knowledge Graphs", *Data Science and Engineering* 5(3), p. 293-316.
67. https://marilenadaquino.github.io/hico/; Marilena Daquino & Francesca Tomasi (2015), "Historical Context Ontology (HiCO): A Conceptual Model for Describing Context Information of Cultural Heritage Objects", in Emmanouel Garoufallou, Richard J. Hartley & Panorea Gaitanou (eds.), *Metadata and Semantics Research*, Cham, Springer International Publishing, p. 424-436.

ABSTRACTS

In this paper, we challenge the vision of creating a giant knowledge graph of reusable research data from multiple projects and platforms, based on Semantic Web technologies and a shared conceptualisation in the form of a commonly used ontology ecosystem, promoted by the *Semantic Data for Humanities and Social Sciences* project (SDHSS). Three main issues are raised: the difficulty of proposing a cross-disciplinary ontology for humanities and social sciences given the different research agendas and constructivist approaches; the difficulty of interlinking the objects present in the different information systems, due to different notions of identity; the challenge of dealing with uncertainty, inconsistencies and contradictions in information sources, when the aim is to produce 'factual' information. First, I propose an epistemological analysis of information, as distinct from data and knowledge, then I develop a conceptualisation of information in humanities and social sciences, based on the OntoClean methodology, and the foundational ontologies DOLCE and Descriptions and Situations, and show how it can ground a cross-disciplinary ontology ecosystem based on the CIDOC CRM and its SDHSS integration. Finally, I apply these principles to a detailed analysis of the information production process and show how the proposed conceptualisation provides possible solutions to the three challenges raised.

Dans cet article, nous mettons au défi le programme promu par le Projet « Semantic Data for Humanities and Social Sciences » (SDHSS) visant la création d'un graphe géant d'information à partir de données de recherche réutilisables en provenance de multiples projets et plateformes, graphe basé sur les technologies du Web sémantique et sur une conceptualisation collaborative en l'espèce d'un écosystème d'ontologies interopérables et réutilisables. Trois questions principales sont abordées : la difficulté de proposer une ontologie interdisciplinaire pour les sciences humaines et sociales, compte tenu des différents agendas de recherche et des différentes approches constructivistes. La difficulté de relier entre eux les objets dans les différents systèmes d'information, en raison des différentes notions d'identité. Le défi de traiter l'incertitude, les incohérences et les contradictions dans les sources d'information, lorsque l'objectif est de produire de l'information « factuelle ». Je propose d'abord une analyse épistémologique de l'information, distincte des données et des connaissances. Je développe ensuite une conceptualisation de l'information en sciences humaines et sociales, basée sur la méthodologie OntoClean et les ontologies fondationnelles DOLCE et Descriptions et Situations, en montrant comment elle peut fonder un écosystème d'ontologies transdisciplinaires basé sur le CRM CIDOC et son intégration SDHSS. Pour finir, j'applique ces principes à une analyse détaillée du processus de production de l'information et montre comment la conceptualisation proposée fournit des solutions aux trois défis en question.

INDEX

Mots-clés: Ontologies transdisciplinaires, épistémologie de la production d'information dans les sciences humaines et sociales, OntoClean, DOLCE et Descriptions and Situations, CIDOC CRM, SDHSS, factoïdes, processus de production d'informations factuelles
Keywords: cross-disciplinary ontology, epistemology of information production in the humanities and social sciences, OntoClean, DOLCE and Descriptions and Situations, CIDOC CRM, SDHS, factoids, process of factual information production

AUTHOR

FRANCESCO BERETTA

Laboratoire de recherche historique Rhône-Alpes, CNRS / Université de Lyon